DOCUMENT RESUME

ED 317 619 TM 014 744

AUTHOR Lunz, Mary E.; And Others
TITLE Test-Retest Consistency of Computer Adaptive Tests.
PUB DATE Apr 90
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability Identification; *Adaptive Testing; *College Students; Comparative Testing; *Computer Assisted Testing; Higher Education; Item Response Theory; Licensing Examinations (Professions); *Mastery Tests; Maximum Likelihood Statistics; Test Length; *Test Reliability
IDENTIFIERS Rasch Model; *Test Retest Reliability

ABSTRACT

This study explores the test-retest consistency of computer adaptive tests of varying lengths. The testing model used was designed as a mastery model to determine whether an examinee's estimated ability level is above or below a pre-established criterion expressed in the metric (logits) of the calibrated item pool scale. The Rasch model was used to calibrate items and estimate person measures. The calibrated item pool contained 726 items. The PROX version of the maximum likelihood method of item selection was used in the adaptive algorithm. The content coverage in the computer adaptive testing was designed to be comparable to the test specifications for the conventional paper-and-pencil certification examination. A total of 765 students from across the nation participated as examinees in this study; a random sample of 162 of these students was placed in the test-retest condition. Examinees took two contiguous tests with the same test specifications, but different items (alternative forms of varying lengths). The ability measures from the test and retest were found to correlate at 0.95 when attenuated for error, demonstrating that differentiation among examinee measures is comparable regardless of the length of the test or the particular subset of items. This finding provides evidence of test-retest consistency of computer adaptive tests. Two data tables are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317619

U.S. DEPARTMENT OF EDUCATION
(Office of Educational Research and Improvement)
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARY E. LUNZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Test-Retest Consistency of Computer Adaptive Tests

Mary E. Lunz

Betty A. Bergstrom

Richard Gershon

American Society of Clinical Pathologists

Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA., April, 1990.

TM017777

BEST COPY AVAILABLE

Test - Retest Consistency of Computer Adaptive Tests

This study explores the test-retest consistency of computer adaptive tests of varying lengths. Examinees took two contiguous tests with the same test specifications but different items (alternate forms of varying lengths). The ability measures from the test and retest were found to correlate at .95 when attenuated for error, demonstrating that differentiation among examinee measures is comparable regardless of the length of the test or the particular subset of items. This provides evidence of the test-retest consistency of computer adaptive tests.

Test - Retest Consistency of Computer Adaptive Tests

One approach to computer administered tests is adaptive testing, a process in which items are selected from calibrated item pools based on the examinees responses to previously administered items. The algorithms for item selection often depend on item response theory (IRT) (Lord and Novick, 1968; Wright and Stone, 1979). The difficulties of the items in the pool are calibrated to a common log-linear scale so that items with difficulties targeted to the ability of a given examinee can be selected. Each examinee takes a unique alternate test form tailored to his individual ability.

Wright (1968 and 1977) argues for item free measurement, which means that, when all items are calibrated on the same scale, equivalent person measures should result regardless of the particular items included on a test form. This is a requirement for computer adaptive testing. Rudner (1989) states that with computer adaptive testing, students take different forms of the same test because all items are calibrated to the same scale. Weiss (1985) found that computer adaptive testing could reduce the length of the test by 50% with no appreciable effect on the psychometric characteristics and improves the efficiency of classification of examinees.

For high-stakes testing such as certification, it is necessary to establish confidence in the psychometric properties of computer adaptive tests prior to implementation of computer adaptive strategies. If it can be shown that pass/fail decisions are valid and reliable regardless of the items presented, then computer adaptive tests can offer many advantages to examinees, such as decreased test length and increased test availability. The

content validity of computer adaptive tests can be supported if the calibrated item pool is comparable in content to current certification examinations, and is designed to cover the field of practice. The same content specifications can be used when the test is administered by computer or paper and pencil (see Lunz and Stahl, 1988 for content validity procedures). The psychometric issue of the test/retest or alternate forms reliability of computer adaptive tests, however, requires careful study.

Examinees take different and variable length tests in adaptive testing rather than identical fixed length tests. This raises the question of whether all examinees will be tested comparably and whether decisions will be equally valid when different but parallel computer adaptive tests are administered. The items included in a computer adaptive test are those which provide the most information about the ability of the examinee. This allows the test to be shorter. Alternate computer adaptive test forms are parallel because they adhere to the same test specifications.

Theoretically, any subset of items selected adaptively from a calibrated item pool should produce a statistically equivalent measure for an examinee of a given ability (Wright, 1979). It follows that two computer adaptive tests administered contiguously to the same examinee should produce comparable measures of that examinee's ability.

The Standards for Educational and Psychological Testing (1985) state that the goal in the study of reliability is to estimate the magnitude of the error of measurement or to estimate the degree of consistency of the scores on two parallel tests. Even very careful matching of item content and difficulty on two tests can not ensure that an examinee will know the answer to a particular item and its matched counterpart, so scores will rarely be

precisely the same. Stanley (1978) defines reliability as the tendency toward consistency from one set of measures to another. Allen and Yen (1979) define reliability as the correlation between observed scores on parallel tests. Reliabilty between parallel computer adaptive tests was addressed in a study by Martin, McBride and Weiss (1983). Scores on two alternate fixed length forms of adaptive tests were found to correlate at .90, attenuated for error, after thirty items were administered.

The purpose of this paper is to report test-retest reliability when alternate parallel forms of computer adaptive tests are administered sequentially to the same examinee. Correlations between tests and comparability of pass/fail decisions provide evidence of test-retest consistency.

Method

The computer adaptive testing model used in this study has the following characteristics. It is designed as a mastery model (Weiss and Kingsbury, 1984) to determine whether a person's estimated ability level is above or below a pre-established criterion expressed in the metric (logits) of the calibrated item pool scale. The Rasch model (Rasch, 1960/1980) was used to calibrate items and estimate person measures. The Rasch model was selected because the sample sizes were not large enough to meet the requirements of 2 and 3 parameter models (Lord, 1983) and there is evidence that the person measures estimated with the Rasch and the 2 and 3 parameter models correlate highly (.90) when tests are administered under a computer adaptive algorithm (Olsen, et al, 1986).

The calibrated item pool contained 726 items. The PROX version of the maximum likelihood method of item selection (Wright and Stone, 1979) was used

in the adaptive algorithm. In addition a content balancing mechanism of the type described by Kingsbury and Zara (1989) was included. The content coverage in the computer adaptive testing was designed to be comparable to the test specifications for the conventional paper and pencil certification examination.

The computer adaptive test administrator program (Gershon, 1989) was designed to minimize examinee waiting time between items. While the examinee considered the item presented, the computer selected two items within the specified content area, one which would yield maximum information should the current item be answered incorrectly and another which would yield maximum information should the current item be answered correctly. The examinee was required to answer the item before another was presented. The opportunity to review or change answers at a later time was not available.

Two test conditions were defined based on a minimum number of items: a 'short test' with a minimum of 50 items and a 'long test' with a minimum of 100 items. Maximum test length varied. The stopping rule required the measured ability of examinees to be 1.3 times the SEM (Wright and Masters, 1982) above or below the pass point before testing stopped. This applies a 90 percent confidence interval (one-tailed test) around the pass/fail decision. If the examinee was a clear pass or fail after completing the minimum number of items (50 or 100), testing stopped. If, however, the examinee measure was not 1.3 SEM above or below that pass point, testing continued until the requirement was achieved or until the examinee answered 240 questions. Test lengths varied to meet the stopping rule. A pass/fail decision was made based on the examinee measure after 240 items were administered whether or not the examinee was a clear pass or fail as defined by the stopping rule.

For certification exams, precision (i.e. low SEM) is desirable for examinees with measures near the pass point. Precision, however, is not as crucial for those examinees who are clearly above or below the pass point early in the test. The stopping rule was designed to increase classification validity by increasing precision near the pass point.

Seven-hundred sixty-five (765) students from across the country participated as examinees in this study, under the auspices of their professional education program. A random sample of 162 of these students was placed in the test-retest condition. Students who participated were serious about the experience since it was part of their review for the certification examination.

The examinees in the test-retest condition took two contiguous computer adaptive tests but were unaware that they were taking two separate tests. The two separate tests had comparable test specifications for content coverage, starting point and stopping rule. Items were presented until the test specifications were completed and a decision about the examinee could be made. The number of items presented in the test and retest was not fixed and often varied on the test and retest. Items presented to an examinee on the first test were marked by the computer so they would not be administered to the same examinee on the retest. Correlation of the ability measures and comparability of the pass/fail decisions from the test and the retest are the basis for determining the alternate forms test-retest reliability of this computer adaptive test.

Results

The correlation between the test and retest for the short test (minimum 50 items) was .95, corrected for attenuation. The correlation between the

test and retest for the long test (minimum 100 items) was .95, corrected for attenuation. Table 1 summarizes the decision consistency on the test and retest. When examinees are clear passes or fails (1.3 SEM above or below the pass/fail point) for both the test and the retest, the same pass/fail decision was made for 98 of the 107 examinees or 92%. This indicates that for most of the examinees, ability measures were comparable on the test and retest regardless of the length of the test or the particular items administered. Only 9 of these 107 examinees changed status and had significantly different measures. Fifty-five examinees were within the error of measure on one or both of the tests so that a clear pass/fail decision using the established stopping rule, could not be made. When a clear decision could not be made, ability measures were within ±.4 logits of the pass/fail point. The same pass/fail decision was made for 37 of these 55 examinees or 67%.

The number of items included on the test and the retest varied more than anticipated. For some examinees the test lengths were the same, for others, the initial test was longer or shorter than the retest or vice versa. The length of the test depended on the relationship of the examinee measure to the pass/fail point. Those examinees who were not a clear pass or fail took longer tests. Examinees in the short test condition took an average of 81 items on the test and 86 items on the retest (range = 50-240). Examinees in the long test condition took an average of 121 items on the test and an average of 130 items on the retest (range = 100-240).

Table 2 presents the summary statistics for test and retest for the short and long tests. The standard deviations and the standard errors of measure were comparable between the test and retest for short and long tests. The person separation reliability specifies the proportion of observed

variance not due to estimation error. This statistic is similar to the KR-20 and indicates the extent of useful variations among examinee measures (Wright and Masters, 1982). The person separation reliabilities were comparable for the test (.77) and the retest (.80) for the short test and identical for the test and retest (.89) for the long test. The SEM is smaller for the long test (.19 long test compared to .26 short test) because the minimum number of items administered was greater.

Discussion

Ability measures from alternate computer adaptive tests correlate well with each other when administered to the same examinees under comparable conditions. Even though the length of the tests and the particular subset of items selected vary, statistically equivalent ability measures are obtained despite the differences in the test forms. For most examinees, the same pass/fail decision is made. Even when examinee measures hover near the pass/fail point, the same decision is made for 67% of the examinees.

When the length of the computer adaptive test increases, the person separation reliability improves slightly as the standard error of measure declines. Psychometricians have argued that "longer is better" because the measurement error decreases as the number of items increases. Computer adaptive testing greatly decreases the range of item difficulties sampled by tailoring to the ability of the examinee. This maximizes the efficiency of each item.

In this study, a 50 item test and retest appeared as 100 items to the examinee. Those examinees with measures within ± 1.3 SEM of the pass point were administered more items in an attempt to make a definitive pass/fail decision. When this failed, the testing stopped after 240 items. Two

consequences may have resulted from this design. (1) Fatigue on the second test may have become a factor for some examinees. Table 2 shows a lower mean measure for the second test in the long test condition. Also, Table 1 shows that when a clear pass/fail decision was not made, more examinees passed the test and failed the retest. (2) The item pool was taxed severely when large numbers of items were administered (ie one examinee answered 417 items from the pool of 726). As the number of unused items in the pool declined, less well targeted items were presented to the examinee, so that more items were needed to meet the required level of precision. Examinees who took 240 items (some of which had to be off target) did so because they never met the requirement of the stopping rule. The testing stopped because the maximum number of items was administered not because a clear pass/fail decision was made. The value of presenting more items may have been out weighted by examinee fatigue and the limitations of the item pool.

Very few research projects have explored test-retest reliability of alternate test forms of computer adaptive tests. The issues of test length and stopping rule address the problem of when testing can reasonably be terminated. They are complex and need further exploration. We hope this study will inspire more research since a basic premise of computer adaptive tests is that alternate, but tailored tests can be used to make comparable decisions.

Table 1

Pass/Fail Consistency
on Test and Retest

CLEAR PASS/FAIL**
Decisions on Test and Retest

		RETEST: Pass	RETEST: Fail		
TEST	Pass	52*	5	57	* 98/107 (92%) same pass/fail decision on test and retest
	Fail	4	46*	50	
		56	51	107	

NO CLEAR PASS/FAIL DECISION REACHED ON THE TEST OR RETEST OR BOTH***

		RETEST: Pass	RETEST: Fail		
TEST	Pass	17*	13	30	*37/55 (67%) same pass/fail decision on test and retest
	Fail	5	20*	25	
		22	33	55	

** Examinee measures are ≥ 1.3 times SEM above or below the pass/fail point for both the test and the retest.

*** Examine measures are < 1.3 times SEM above or below the pass/fail point on the test, retest or both tests.

Table 2

Summary Statistics for Test and Retest

Short Test (minimum 50 items)	Test	Retest	Correlation***
N examinees	105	105	
N items	50-240	50-240	
Mean number of items	81	86	
Measure			
Mean	.20	.19	
SD	.55	.56	
Standard Error of Measure			
Mean	.26	.25	
MSqE*	.07	.07	
Reliability**	.77	.78	.95

Long Test (minimum of 100 items)	Test	Retest	
N examinees	57	57	
N items	100-240	100-240	
Mean number of items	121	130	
Measure			
Mean	.12	.03	
SD	.59	.59	
Standard Error of Measure			
Mean	.19	.19	
MSqE*	.04	.04	
Reliability**	.89	.89	.95

Combined (all test lengths)	Test	Retest	
N examinees	162	162	
N items	50-240	50-240	
Measure			
Mean	.17	.13	
SD	.57	.57	
Standard Error of Measure			
Mean	.24	.23	
MSqE*	.06	.06	
Reliability**	.82	.82	.95

*MSqE = mean squared error; mean of the squared error associated with each person measure; (Length of Test/Right x Wrong)½

**Reliability (reliability of person separation) = $(SD^2\text{-}MSqE)/SD^2$

***Correction for attenuation = $r_{xy}/(r_{xx}r_{yy})^{½}$

References

Allen, M.J. and Yen, W.M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.

Gershon, R.C., (1989). CAT Administrator (Computer Program) Chicago: Micro Connections.

Kingsbury, G.G. and Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2:4, 359-375.

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Score*. Reading, Mass: Addison-Wesley.

Lord, F.M. (1983). Small N Justifies Rasch Methods. In D. Weiss (Ed.), *New Horizons in Testing*. New York: Academic Press.

Lunz, M.E. and Stahl, J.A. (1989). Content validity revisited: transforming job analysis data into test specifications. *Evaluation and the Health Professional*, 12, 192-206.

Martin, J.T. McBride, J.R., Weiss, D.J. *Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population (1983)*. Research Report 83-1, Minneapolis, MN: University of Minnesota.

Olsen, J.B., Maynes, D.D., Slawson, D. and Hoy, K. (1986). *Comparison and Equating of Paper-administered, Computer-assisted and Computerized Adaptive Tests of achievement*. Paper presented at the annual meeting of the American Educational Research Association.

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Rudner, L.M. (1989). *Educational Measurement Issues and Practice*, 8:4, 25-26 Notes From Eric/TM.

Standards for Educational and Psychological testing (1985). American Psychological Association, Washington, D.C.

Stanley, J.C. (1978). "Reliability" In Thorndike, R. (ed.). *Educational Measurement*. Washington, D.C.: American Council on Education.

Weiss, D.J. (1985). *Final Report: Computerized Adaptive Measurement of Achievement and Ability*, Project NR150-433, N00014-79-C-0172, Minneapolis, MN: University of Minnesota.

Weiss, D.J. and Kingsbury, G.G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measurement*, 21:4,361-375.

Wright, B.D. (1968). Sample free calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service.

Wright, B.D. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, 14, 97-116.

Wright, B.D. and Masters, G. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Wright, B.D. and Stone, M. (1979). *Best Test Design*. Chicago, IL: MESA Press.